|  |
| --- |
|  |
| SMDM PROJECT REPORT |
| DSBA |
|  |
| **Shripad Anwekar / PGPDSBA Online Mar\_A2021** |
| **5/9/2021** |

|  |
| --- |
|  |

Contents

[1. Problem 1 3](#_Toc71480413)

[1.1. Use methods of descriptive statistics to summarize data. 3](#_Toc71480414)

[1.1.1. Data Summary 3](#_Toc71480415)

[1.1.1. Which Region and which Channel spent the most? 4](#_Toc71480416)

[1.1.2. Which Region and which Channel spent the least? 4](#_Toc71480417)

[1.2. There are 6 varieties of s that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer 5](#_Toc71480418)

[1.3. On the basis of a descriptive measure of variability, which shows the most inconsistent behaviour? Which show the least inconsistent behaviour? 12](#_Toc71480419)

[1.4. Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments. 13](#_Toc71480420)

[1.5. On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective 16](#_Toc71480421)

[2. Problem 2 17](#_Toc71480422)

[2.1. For this data, construct the following contingency tables (Keep Gender as row variable) 17](#_Toc71480423)

[2.1.1. Gender and Major 17](#_Toc71480424)

[2.1.2. Gender and Grad Intention 17](#_Toc71480425)

[2.1.3. Gender and Employment 17](#_Toc71480426)

[2.1.4. Gender and Computer 18](#_Toc71480427)

[2.2. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 19](#_Toc71480428)

[2.2.1. What is the probability that a randomly selected CMSU student will be male? 19](#_Toc71480429)

[2.2.2. What is the probability that a randomly selected CMSU student will be female? 19](#_Toc71480430)

[2.3. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 20](#_Toc71480431)

[2.3.1. Find the conditional probability of different majors among the male students in CMSU. 20](#_Toc71480432)

[2.3.2. Find the conditional probability of different majors among the female students of CMSU. 20](#_Toc71480433)

[2.4. Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question: 21](#_Toc71480434)

[2.4.1. Find the probability that a randomly chosen student is a male and intends to graduate. 21](#_Toc71480435)

[2.4.2. Find the probability that a randomly selected student is a female and does NOT have a laptop. 21](#_Toc71480436)

[2.5. Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question: 21](#_Toc71480437)

[2.5.1. Find the probability that a randomly chosen student is a male or has full-time employment? 21](#_Toc71480438)

[2.5.2. Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management. 21](#_Toc71480439)

[2.6. Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events? 22](#_Toc71480440)

[2.7. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. 23](#_Toc71480441)

[2.7.1. If a student is chosen randomly, what is the probability that his/her GPA is less than 3? 23](#_Toc71480442)

[2.7.2. Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more. 23](#_Toc71480443)

[2.8. Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions for this whole Problem 2. 24](#_Toc71480444)

[2.8.1. For each of them comment whether they follow a normal distribution 24](#_Toc71480445)

[2.8.2. Write a note summarizing your conclusions for this whole Problem 2. 33](#_Toc71480446)

[3. Problem 3 34](#_Toc71480447)

[3.1. Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps. 34](#_Toc71480448)

[3.2. Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed? 36](#_Toc71480449)

# Problem 1

A wholesale distributor operating in different regions of Portugal has information on annual spending of several s in their stores across different regions and channels. The data consists of 440 large retailers’ annual spending on 6 different varieties of products in 3 different regions (Lisbon, Oporto, Other) and across different sales channel (Hotel, Retail).

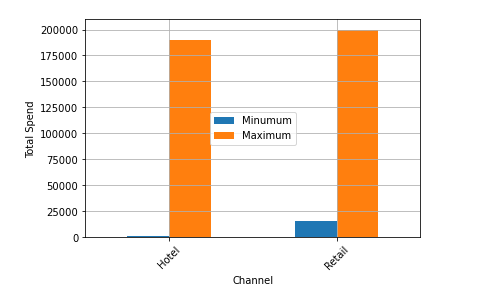
## Use methods of descriptive statistics to summarize data.

### ****Data Summary****

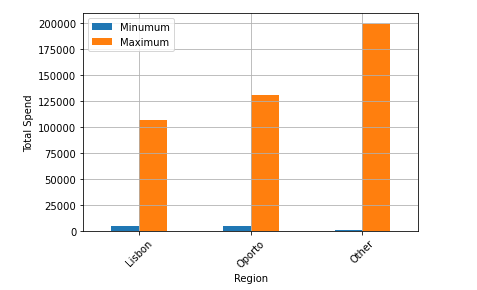
* There are 440 rows and 9 columns in the Datasets
* There are no null value present in the data Set
* Buyer/Spender is the **identification No** of Buyer/Spender
* Channel is the category of Sales Channel – **Hotel or Retail**
* Region is the division that covers the sell the designated area – **Oporto, Lisbon and Other**
* There are 6 items sold through these channels across these regions - **Fresh, Milk, Grocery, Frozen, Detergents\_Paper and Delicatessen**
* Each spender / buyer spends on each item
* **Total spend** is Maximum on ‘Fresh’ and Minimum on ‘Delicatessen’
* **Average spend** is Maximum on ‘Fresh’ and Minimum on ‘Delicatessen’

Following Steps are followed to identify the required information

* ‘Total Spend’ is calculated by adding the spend on each by the Spender.
* ‘Bar Plot’ is then plotted to visually present the data for Comparison.
* Channel Wise Maximum and Minimum Spend are plotted in Plot 1.1.1
* Region Wise Maximum and Minimum Spend are plotted in Plot 1.1.2



***Plot 1.1.1***



***Plot 1.1.2***

In order to ascertain the Plots; descriptive statistics are reviewed, please refer to **Table 1.1.1** for

Numbers visualised on the plots

|  |  |  |
| --- | --- | --- |
| Spend | Minimum Spend | Maximum Spend |
| Amount | 904 | 199891 |
| Channel | Hotel | Retail |
| Region | Other | Other |

***Table 1.1.1***

## 

### ****Which Region and which Channel spent the most?****

Please refer to ***Table 1.1.1***; Maximum of ‘Total Spend’ of **199891** is through Channel ‘Retail’ and by Region 'Other '.

### ****Which Region and which Channel spent the least?****

Please refer to ***Table 1.1.1***; Minimum of ‘Total Spend’ of **904**  is through Channel ‘Hotel’ and by Region 'Other '.

## There are 6 varieties of s that are considered. Describe and comment/explain all the varieties across Region and Channel? Provide a detailed justification for your answer

#### There are 6 varieties of s Fresh, Milk, Grocery, Frozen, Detergents\_Paper and Delicatessen. They are sold across three regions Lisbon, Oporto and Other. The selling is done through Channels; Hotel and Retail.

Across the Region Analysis

1) In this exercise the data set containing 440 rows and 9 Columns is analysed

2) None of the values are null.

3) Data is analysed for the count, central tendency, 5 number summaries, consistency (CV) and Skewness of the distribution

4) Please refer to ***Table 1.2.1 for*** Spender Distribution

|  |  |
| --- | --- |
| Region | Spender Count |
| Oporto | 47 |
| Lisbon | 77 |
| Other | 316 |

***Table 1.2.1***

5) ‘Average Sale’ performance of Items based on region is summarised in ‘***Table 1.2.2’***

|  |  |  |  |
| --- | --- | --- | --- |
|  | Performance by Region | | |
| **Item** | Good | Average | Need to Improve |
| Fresh | Other | Lisbon | Oporto |
| Milk | Other | Lisbon | Oporto |
| Grocery | Oporto | Other | Lisbon |
| Frozen | Oporto | Other | Lisbon |
| Detergent\_Paper | Oporto | Other | Lisbon |
| Delicatessen | Other | Lisbon | Oporto |

***Table 1.2.2***

Further Data is analysed for the count, central tendency, 5 number summaries, consistency (CV) and skewness of the distribution is summarised in **Table 1.2.3**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Parameter | Region | Item | Spend | Total Spenders | Spenders Impacted | R/G/Y | Comments |
| Minimum Average Spend | Oporto | Delicatessen | 1159.7 | 47 | 32 |  | This need to be reviewed for strategy and rates . **32** spenders who are spending less than minimum of average have improvement possibilities Also Please refer to ***Table 1.2.2*** for performance /  Recommendation where combination of Oporto / Delicatessen Needs review, which aligns with this analysis. |
| Maximum average spend | Other | Fresh | 12533.47 | 316 | 203 |  | This Item / region combination seem to be doing well however **203** prospects have improvement possibilities |
| Minimum Standard Deviation | Oporto | Delicatessen | 1050.74 | 47 | 27 |  | As Mentioned against parameter “Minimum Average Spend “  Please refer to ***Table 1.2.2*** for performance /  Recommendation where combination of Oporto / Delicatessen Needs review, which aligns with this analysis. |
| Maximum Standard Deviation | Other | Fresh | 13389.21 | 316 | 210 |  | This aligns with “Maximum Average Spend “  For combination of ‘Region/ Item’, Also please refer to ***Table 1.2.2*** for performance / recommendation |
| Minimum Spend | Other | Grocery | 3 | 440 | 1 |  | This one is Outlier; Needs further analysis. |
| 25% of spend (Minimum) | Other | Detergents\_Paper | 251.25 | 316 | 79 |  | Minimum spend on any Item by lowest 25% of spenders. Please refer to ***Table 1.2.2*** for performance /  Recommendation where combination of Other / Detergents\_Paper Needs review, which aligns with this analysis. |
| 25% of spend (Maximum) | Other | Fresh | 3350.75 | 316 | 237 |  | Maximum spend on any Item by lowest 25% of spenders.  Also please refer to ***Table 1.2.2*** for performance / recommendation |
| 50% of spend (Minimum) | Lisbon | Detergents\_Paper | 737.00 | 77 | 38 |  | Minimum spend on any Item by lowest 50% of spenders. Please refer to ***Table 1.2.2*** for performance /  Recommendation where combination of Lisbon / Detergents\_Paper Needs review, which aligns with this analysis. |
| 50% of spend (Maximum) | Other | Fresh | 8752.50 | 316 | 158 |  | Maximum spend on any Item by lowest 50% of spenders.  Also please refer to ***Table 1.2.2*** for performance /  Recommendation where combination of Other / fresh is doing good which aligns with this analysis. |
| 75% of spend (Minimum) | Oporto | Delicatessen | 1538.50 | 47 | 35 |  | Minimum spend on any product by lowest 75% of spenders. Please refer to ***Table 1.2.2*** for performance /  Recommendation where combination of Oporto / Delicatessen Needs review, which aligns with this analysis. |
| 75% of spend (Maximum) | Other | Fresh | 17406.50 | 316 | 237 |  | Maximum spend on any Item by lowest 75% of spenders.  ***Table 1.2.2*** for performance /  Recommendation where combination of Other / fresh is doing good which aligns with this analysis. |
| Maximum Spend | Other | Fresh | 112151 | 440 | 1 |  | Maximum spend on any Item by spenders. Please refer  ***Table 1.2.2*** for performance /  Recommendation where combination of Other / fresh is doing good which aligns with this analysis. |
| CV (Most Consistent) | Oporto | Fresh | 0.85 |  |  |  | Please refer  ***Table 1.2.2*** for performance /  Recommendation where combination of Oporto / fresh is not doing so good. However it is showing a consistent sale. |
| CV (least Consistent) | Oporto | Frozen | 2.26 |  |  |  | Please refer  ***Table 1.2.2*** for performance /  Recommendation where combination of Oporto / Frozen is doing really good overall. However it is showing least consistent sale. |
| ['Max - 50%'] > ['Q2 - Min'] ) |  | | | | |  | This is calculated for Each Item for Each region; All the values are positive which makes it evident that the individual spread is not Normal but it is Right Skewed.  There is opportunity to improve sell based on the recommendations in T***able 1.2.2*** which will shift the spread to right making it normal as well as improving the sell. |
| ['Max - 75%'] > ['Q1 - Min'] ) |  |
| ['Q3 - Q2'] > ['Q2 - Q1'] ) |  |

***Table 1.2.3***

Across the Channels Analysis

1) In this exercise the data set containing 440 rows and 9 Columns is analysed

2) None of the values are null.

3) Data is analysed for the count, central tendency, 5 number summaries, consistency (CV) and Skewness of the distribution

4) Spender Distribution please refer to ***Table 1.2.4***

|  |  |
| --- | --- |
| Channel | Spender Count |
| Hotel | 298 |
| Retail | 142 |

***Table 1.2.4***

5) ‘Average Sale’ performance of Item based on Channel is summarised in ‘***Table 1.2.5’***

|  |  |  |
| --- | --- | --- |
|  | Performance by Channel | |
| Item | Good | Need to Improve |
| Fresh | Hotel | Retail |
| Milk | Retail | Hotel |
| Grocery | Retail | Hotel |
| Frozen | Hotel | Retail |
| Detergent\_Paper | Retail | Hotel |
| Delicatessen | Retail | Hotel |

***Table 1.2.5***

Further Data is analysed for the count, central tendency, 5 number summaries, consistency (CV) and skewness of the distribution is summarised in **Table 1.2.6**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Parameter | Channel | Item | Spend | Total Spenders | Spenders Impacted | R/G/Y | Comments |
| Minimum Average Spend | Hotel | Detergents\_Paper | 790.56 | 298 | 212 |  | This need to be reviewed for strategy and rates . **212** spenders who are spending less than minimum of average have improvement possibilities Also Please refer to ***Table* 1.2.5** for performance /  Recommendation where combination of Hotel / Detergents\_Paper Needs review, which aligns with this analysis. |
| Maximum average spend | Retail | Grocery | 16322.85 | 142 | 93 |  | This Item / region combination seem to be doing well however **93** prospects have improvement possibilities |
| Minimum Standard Deviation | Hotel | Detergents\_Paper | 1104.09 | 298 | 239 |  | This is interesting point because Channel (Hotel) on Item (Detergents\_Paper) has the minimum Average Spend across .  Hence lowest Standard deviation for this is not favourable from Sales standpoint.  As Mentioned against parameter “Minimum Average Spend “  Please refer to ***Table* 1.2.5**for performance /  Recommendation where combination of Hotel / Detergents\_Paper Needs review, which aligns with this analysis. |
| Maximum Standard Deviation | Hotel | Fresh | 13831.69 | 298 | 193 |  | This is favourable point because Channel ( Hotel) on Item (Fresh) has the maximum Average Spend across hence Highest Standard deviation for this is favourable from Sales standpoint.  This aligns with “Maximum Average Spend “ For combination of ‘Channel/ Item, Also please refer to ***Table* 1.2.5**for performance / recommendation . |
| Minimum Spend | Hotel | Fresh | 3 | 440 | 1 |  | This one is Outlier; Needs further analysis. |
| 25% of spend (Minimum) | Hotel | Detergents\_Paper | 183.25 | 298 | 75 |  | Minimum spend on any Item by lowest 25% of spenders. Please refer to ***Table* 1.2.5**for performance /  Recommendation where combination of Hotel / Detergents\_Paper Needs review, which aligns with this analysis. |
| 25% of spend (Maximum) | Retail | Grocery | 9245.25 | 142 | 36 |  | Maximum spend on any Item by lowest 25% of spenders.  Also please refer to ***Table* 1.2.5**for performance / recommendation where combination of Retail / Grocery shown as doing well, which aligns with this analysis. |
| 50% of spend (Minimum) | Hotel | Detergents\_Paper | 385.50 | 298 | 148 |  | Minimum spend on any Item by lowest 50% of spenders. Please refer to ***Table* 1.2.5** for performance /  Recommendation where combination of Hotel / Detergents\_Paper Needs review, which aligns with this analysis. |
| 50% of spend (Maximum) | Retail | Grocery | 12390 | 142 | 71 |  | Maximum spend on any Item by lowest 50% of spenders.  Also please refer to ***Table* 1.2.5** for performance /  Recommendation where combination of Retail / Grocery is doing good which aligns with this analysis. |
| 75% of spend (Minimum) | Hotel | Detergents\_Paper | 899.50 | 298 | 223 |  | Minimum spend on any Item by lowest 75% of spenders. Please refer to ***Table* 1.2.5** for performance /  Recommendation where combination of Hotel / Detergents\_Paper Needs review, which aligns with this analysis. |
| 75% of spend (Maximum) | Retail | Grocery | 20183.50 | 142 | 106 |  | Maximum spend on any Item by lowest 75% of spenders.  ***Table* 1.2.5** for performance /  Recommendation where combination of Retail / Grocery is doing good which aligns with this analysis. |
| Maximum Spend | Hotel | Detergents\_Paper | 112151.00 | 440 | 1 |  | Maximum spend on any Item by spenders. Please refer  ***Table* 1.2.5** for performance /  Recommendation where combination of Hotel / Detergents\_Paper is doing good which aligns with this analysis. |
| CV (Most Consistent) | Retail | Grocery | 0.75 |  |  |  | Please refer  ***Table* 1.2.5** for performance /  Recommendation  has been mention "Good" ; together with the good Sale; this Channel is also less volatile or risky as the variation around mean is minimal across the Items and channel. |
| CV (least Consistent) | Hotel | Delicatessen | 2.22 |  |  |  | Please refer  ***Table* 1.2.5** for performance /  Recommendation as  "Need Improvement" for the Sell; though the Sell Not so good for this Channel however it is also shown more volatile or risky as the variation around mean is maximum across the Items and region.  This need to be reviewed |
| ['Max - 50%'] > ['Q2 - Min'] ) |  | | | | |  | This is calculated for Each Item for Each Channel; All the values are positive which makes it evident that the individual spread is not Normal but it is Right Skewed.  There is opportunity to improve sell based on the recommendations in ***Table* 1.2.5** which will shift the spread to right making it normal as well as improving the sell. |
| ['Max - 75%'] > ['Q1 - Min'] ) |  |
| ['Q3 - Q2'] > ['Q2 - Q1'] ) |  |

***Table 1.2.6***

## On the basis of a descriptive measure of variability, which shows the most inconsistent behaviour? Which show the least inconsistent behaviour?

Based on the Spend across individual products on 440 items Coefficient of Variation has been calculated which is summarised in ***Table 1.3.1***

|  | **CV (Std/ Mean)** |
| --- | --- |
| **Delicatessen** | 1.85 |
| **Detergents\_Paper** | 1.65 |
| **Frozen** | 1.58 |
| **Milk** | 1.27 |
| **Grocery** | 1.20 |
| **Fresh** | 1.05 |

***Table 1.3.1***

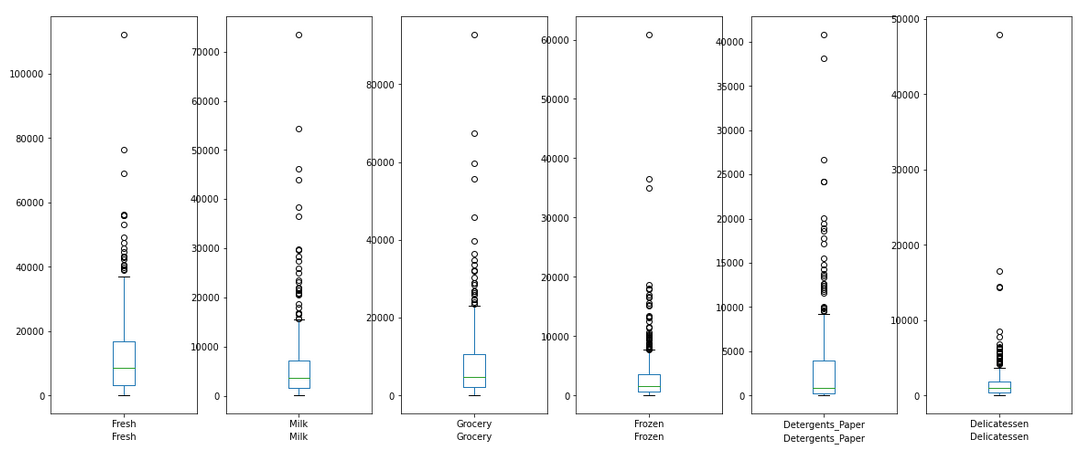
##### Coefficient of Variation is Maximum for Item **"Delicatessen"** making it most volatile/ Risky and Inconsistent Item with Coefficient of variation as - **1.85**

##### Coefficient of Variation is Minimum for Item **"Fresh"** making it most reliable and consistent Item with Coefficient of variation as - **1.05**

## Are there any outliers in the data? Back up your answer with a suitable plot/technique with the help of detailed comments.

* Box Plots for each items as shown in **‘Plot 1.4.1’** clearly shows presence of ‘Outliers’.
* Further Analysis on the quantum of outliers across in summarized in table **‘Table 1.4.1’**
* List of outliers across in summarized in table **‘Table 1.4.2’**

As next step, Data in tables in section 1.4 should be taken for further investigation to understand the correctness of these outliers; if these are correct outliers then what can be done to eliminate the negative ones? Or what can be learnt from the Positive ones?

***Plot 1.4.1***

Further Analysis to identify the Itemized count / Proportion of Outliers is shown in **Table 1.4.1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Item | Outlier Count | Outlier % | Outside( Minimum) | Outside( Maximum) |
| Fresh | 20 | 4.55 % | -17603.0 | 37657.0 |
| Milk | 28 | 6.36 % | -6964.75 | 15693.25 |
| Grocery | 24 | 5.45 % | -10620.75 | 23437.25 |
| Frozen | 42 | 9.55 % | -3488.0 | 7788.0 |
| Detergents\_Paper | 30 | 6.82 % | -5261.25 | 9452.75 |
| Delicatessen | 27 | 6.14 % | -1713.5 | 3942.5 |

**Table 1.4.1**

| **Sr. No** | **Fresh** | **Milk** | | **Grocery** | | **Frozen** | | **Detergents\_Paper** | **Delicatessen** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 38793.00 | 15726.00 | 23596.00 | | 7849.00 | | 9529.00 | | 4100.00 |
| **2** | 39228.00 | 15729.00 | 23998.00 | | 7888.00 | | 9606.00 | | 4334.00 |
| **3** | 39679.00 | 16599.00 | 24708.00 | | 8132.00 | | 9836.00 | | 4365.00 |
| **4** | 40254.00 | 16687.00 | 24773.00 | | 8164.00 | | 9959.00 | | 4430.00 |
| **5** | 40721.00 | 16729.00 | 25957.00 | | 8170.00 | | 10069.00 | | 4478.00 |
| **6** | 42312.00 | 16784.00 | 26316.00 | | 8195.00 | | 11577.00 | | 4626.00 |
| **7** | 42786.00 | 17972.00 | 26839.00 | | 8321.00 | | 11783.00 | | 4829.00 |
| **8** | 43088.00 | 18664.00 | 26866.00 | | 8366.00 | | 12034.00 | | 4985.00 |
| **9** | 43265.00 | 20484.00 | 26870.00 | | 8425.00 | | 12218.00 | | 5120.00 |
| **10** | 44466.00 | 20655.00 | 28540.00 | | 8620.00 | | 12408.00 | | 5121.00 |
| **11** | 45640.00 | 20959.00 | 28921.00 | | 8692.00 | | 12420.00 | | 5130.00 |
| **12** | 47493.00 | 21412.00 | 28986.00 | | 8693.00 | | 12591.00 | | 5137.00 |
| **13** | 49063.00 | 21858.00 | 30243.00 | | 8853.00 | | 12638.00 | | 5185.00 |
| **14** | 53205.00 | 22044.00 | 32034.00 | | 8872.00 | | 13308.00 | | 5206.00 |
| **15** | 56082.00 | 23133.00 | 32114.00 | | 9408.00 | | 13583.00 | | 5609.00 |
| **16** | 56083.00 | 23527.00 | 33586.00 | | 9510.00 | | 13726.00 | | 5778.00 |
| **17** | 56159.00 | 25071.00 | 34792.00 | | 9584.00 | | 14235.00 | | 5864.00 |
| **18** | 68951.00 | 25862.00 | 36486.00 | | 9735.00 | | 14841.00 | | 6250.00 |
| **19** | 76237.00 | 27472.00 | 39694.00 | | 9806.00 | | 15469.00 | | 6372.00 |
| **20** | 112151.00 | 28326.00 | 45828.00 | | 9927.00 | | 17120.00 | | 6465.00 |
| **21** | nan | 29627.00 | 55571.00 | | 10002.00 | | 17740.00 | | 6854.00 |
| **22** | nan | 29892.00 | 59598.00 | | 10155.00 | | 18594.00 | | 7844.00 |
| **23** | nan | 36423.00 | 67298.00 | | 10303.00 | | 18906.00 | | 8550.00 |
| **24** | nan | 38369.00 | 92780.00 | | 10643.00 | | 19410.00 | | 14351.00 |
| **25** | nan | 43950.00 | nan | | 11422.00 | | 20070.00 | | 14472.00 |
| **26** | nan | 46197.00 | nan | | 11559.00 | | 24171.00 | | 16523.00 |
| **27** | nan | 54259.00 | nan | | 12569.00 | | 24231.00 | | 47943.00 |
| **28** | nan | 73498.00 | nan | | 13135.00 | | 26701.00 | | nan |
| **29** | nan | nan | nan | | 13223.00 | | 38102.00 | | nan |
| **30** | nan | nan | nan | | 13486.00 | | 40827.00 | | nan |
| **31** | nan | nan | nan | | 15082.00 | | nan | | nan |
| **32** | nan | nan | nan | | 15348.00 | | nan | | nan |
| **33** | nan | nan | nan | | 15601.00 | | nan | | nan |
| **34** | nan | nan | nan | | 16538.00 | | nan | | nan |
| **35** | nan | nan | nan | | 16745.00 | | nan | | nan |
| **36** | nan | nan | nan | | 16919.00 | | nan | | nan |
| **37** | nan | nan | nan | | 17866.00 | | nan | | nan |
| **38** | nan | nan | nan | | 18028.00 | | nan | | nan |
| **39** | nan | nan | nan | | 18711.00 | | nan | | nan |
| **40** | nan | nan | nan | | 35009.00 | | nan | | nan |
| **41** | nan | nan | nan | | 36534.00 | | nan | | nan |
| **42** | nan | nan | nan | | 60869.00 | | nan | | nan |

**Table 1.4.2**

## On the basis of your analysis, what are your recommendations for the business? How can your analysis help the business to solve its problem? Answer from the business perspective

* Identified Regions and Channels with opportunity to improve sell are shown in ‘***Table 1.2.5’*** And ‘***Table 1.2.5’***
* Further Data is analysed for the count, central tendency, 5 number summaries, consistency (CV) and skewness outcome of this analysis is detailed in **‘Table 1.2.3’ and ‘Table 1.2.6’** respectively for **‘Regions’ and ‘Channels’**
* Outcome of analysis is shown in column (R/G/Y), categorised to distinguish the expected actions.
* Actions on Red will turn Yellow in to green on its own as Yellow parameters are by product of Core (5 Number) summary.
* **‘Spenders Impacted’** column in **‘Table 1.2.3’ and ‘Table 1.2.6’** shows the spender count that we have to target improve the sell by reviewing our offerings, client expectations , commercials, strategy.
* Current spread for all the items across regions / channels is **right skewed** i.e. most spend is less than the average. Business Focus on R (Red) categorized findings will shift the spread to right turning the sell Higher.
* **Table** **1.3.1** shows the degrees on inconsistency which shows the fluctuating client behaviour or may be some gap in supply chain. This need to be reviewed to improve.
* Outliers from table **Table 1.4.1 and Table 1.4.2** to be reviewed for correctness.
* This Analysis should be redone after removing the unexpected / error some outliers.

# Problem 2

The Student News Service at Clear Mountain State University (CMSU) has decided to gather data about the undergraduate students that attend CMSU. CMSU creates and distributes a survey of 14 questions and receives responses from 62 undergraduates (stored in the *Survey* data set).

## For this data, construct the following contingency tables (Keep Gender as row variable)

### ****Gender and Major****

Contingency table on data frame ‘CMSU’ between columns ‘Gender’ and ‘ Major’ is outlined in **Table 2.1.1**

| **Major** | **Accounting** | **CIS** | **Economics /Finance** | **International Business** | **Management** | **Other** | **Retailing /Marketing** | **Undecided** | **Total** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |  |  |  |  |  |
| **Female** | 3 | 3 | 7 | 4 | 4 | 3 | 9 | 0 | 33 |
| **Male** | 4 | 1 | 4 | 2 | 6 | 4 | 5 | 3 | 29 |
| **Total** | 7 | 4 | 11 | 6 | 10 | 7 | 14 | 3 | 62 |

**Table 2.1.1**

### ****Gender and Grad Intention****

Contingency table on data frame ‘CMSU’ between columns ‘Gender’ and ‘Grad Intention’ is outlined in **Table 2.1.2**

| **Grad Intention** | **No** | **Undecided** | **Yes** | **Total** |
| --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |
| **Female** | 9 | 13 | 11 | 33 |
| **Male** | 3 | 9 | 17 | 29 |
| **Total** | 12 | 22 | 28 | 62 |

**Table 2.1.2**

### ****Gender and Employment****

Contingency table on data frame ‘CMSU’ between columns ‘Gender’ and ‘Employment’ is outlined in **Table 2.1.3**

| **Employment** | **Full-Time** | **Part-Time** | **Unemployed** | **Total** |
| --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |
| **Female** | 3 | 24 | 6 | 33 |
| **Male** | 7 | 19 | 3 | 29 |
| **Total** | 10 | 43 | 9 | 62 |

**Table 2.1.3**

### ****Gender and Computer****

Contingency table on data frame ‘CMSU’ between columns ‘Gender’ and ‘Computer’is outlined in **Table 2.1.4**

| **Computer** | **Desktop** | **Laptop** | **Tablet** | **Total** |
| --- | --- | --- | --- | --- |
| **Gender** |  |  |  |  |
| **Female** | 2 | 29 | 2 | 33 |
| **Male** | 3 | 26 | 0 | 29 |
| **Total** | 5 | 55 | 2 | 62 |

**Table 2.1.4**

## Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

### ****What is the probability that a randomly selected CMSU student will be male?****

###### Male Count is: 29

###### Female Count is: 33

###### Probability that a randomly selected CMSU student will be a MALE is: Count of Male / Total Students =29/62 = **46.77 %**

### ****What is the probability that a randomly selected CMSU student will be female?****

###### Male Count is: 29

###### Female Count is: 33

###### Probability that a randomly selected CMSU student will be a FMALE is: Count of Female / Total Students = 33/62 = **53.23 %**

## Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

### ****Find the conditional probability of different majors among the male students in CMSU.****

#### Conditional probability of different majors among the male students in CMSU = P (Major | male) = P (Major ∩ male)/ P(male)

**Male P (Major ∩ male)/ P(male**)

Accounting 4 0.14

CIS 1 0.03

Economics/Finance 4 0.14

International Business 2 0.07

Management 6 0.21

Other 4 0.14

Retailing/Marketing 5 0.17

Undecided 3 0.10

### ****Find the conditional probability of different majors among the female students of CMSU.****

#### Conditional probability of different majors among the female students of CMSU = P (Major ∩ female)/ P(female))

**Female P (Major ∩ female)/ P(female))**

Accounting 3 0.09

CIS 3 0.09

onomics/Finance 7 0.21

International Business 4 0.12

Management 4 0.12

Other 3 0.09

Retailing/Marketing 9 0.27

Undecided 0 0.00

## Assume that the sample is a representative of the population of CMSU. Based on the data, answer the following question:

### ****Find the probability that a randomly chosen student is a male and intends to graduate.****

##### Probability That a randomly chosen student is a male and intends to graduate is -

##### P(Grad Intention ∩ Male) = P (Grad Intention| Male) x P (male) = 17/29 \* 29/62 = ***0.27***

### ****Find the probability that a randomly selected student is a female and does NOT have a laptop.****

##### Probability that a randomly selected student is a female and does NOT have a laptop -

##### P(other than Laptop ∩ Female) = (1 - P (Laptop| Female)) x P (Female) = (1- 29/33 )\* 33/62 = ***0.0645***

## Assume that the sample is representative of the population of CMSU. Based on the data, answer the following question:

### ****Find the probability that a randomly chosen student is a male or has full-time employment?****

##### Probability that a randomly chosen student is a male or has full-time employment-

##### P(Male U full-time)= P(Male) + P(Full-Time) -P(Male ∩ Full-Time) = (29/62 + 10/62 - 7/62 ) = ***0.516***

### ****Find the conditional probability that given a female student is randomly chosen, she is majoring in international business or management.****

##### Probability that given a female student is randomly chosen, she is majoring in international business or management.-

##### P(International Business U Management) / Female = P(International Business/Female) + P(Management/Female) = 4/33 + 4/33= ***0.242***

## Construct a contingency table of Gender and Intent to Graduate at 2 levels (Yes/No). The Undecided students are not considered now and the table is a 2x2 table. Do you think the graduate intention and being female are independent events?

Contingency table **including** ‘Undecided’ Grad Intention -

| **Grad Intention** | **No** | **Undecided** | **Yes** |
| --- | --- | --- | --- |
| **Gender** |  |  |  |
| **Female** | 9 | 13 | 11 |
| **Male** | 3 | 9 | 17 |

**Table - 2.6.1**

Contingency table **excluding** ‘Undecided’ Grad Intention -

| **Grad Intention** | **No** | **Yes** |
| --- | --- | --- |
| **Gender** |  |  |
| **Female** | 9 | 11 |
| **Male** | 3 | 17 |

**Table - 2.6.2**

##### Probability that a Female is selected and she intends to graduate while **'undecided'** category considered was ***0.18***

##### Probability that a Female is selected and she intends to Graduate Without **'undecided'** considered increased to ***0.28***

##### The graduate intention and being female are not independent events which is evident from the as the Probability in both the above scenario **differs**.

## Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages.

### ****If a student is chosen randomly, what is the probability that his/her GPA is less than 3?****

##### Count of Students with GPA <3 is 17

##### Total Count of Students = 62

##### If a student is chosen randomly, then the probability that his/her GPA is less than 3 is ***17/62*** = ***27.419%***

### ****Find the conditional probability that a randomly selected male earns 50 or more. Find the conditional probability that a randomly selected female earns 50 or more.****

##### Conditional probability that a randomly selected Male earns 50 or more.

##### Count of Male with salary >= 50 = 14

##### Count of Male = 29

##### Conditional probability that a randomly selected male earns 50 or more is ***0.483***

##### Conditional probability that a randomly selected Female earns 50 or more.

##### Count of Male with salary >= 50 = 18

##### Count of Male = 33

##### Conditional probability that a randomly selected male earns 50 or more is ***0.545***

## Note that there are four numerical (continuous) variables in the data set, GPA, Salary, Spending, and Text Messages. For each of them comment whether they follow a normal distribution. Write a note summarizing your conclusions for this whole Problem 2.

### For each of them comment whether they follow a normal distribution

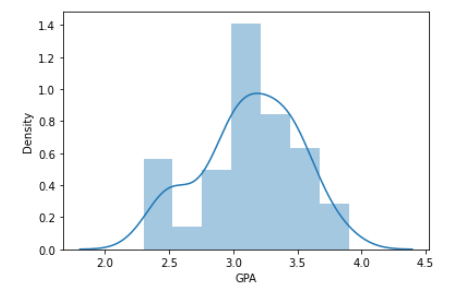
For the variables in question we have analysed the given data against 5 techniques / methods to check if they follow the normal distribution. Outcome of each technique / method together with net conclusion is summarised in **Table 2.8.5**.

The 5 checks applied are as follows –

1. **HISTOGRAM**
2. **5 Number Summaries**
3. **Boxplot**
4. **Empirical Rule**
5. **Mean, Median and Mode Review**

Variable ‘GPA’ Analysis

* **HISTOGRAM (‘GPA’)**



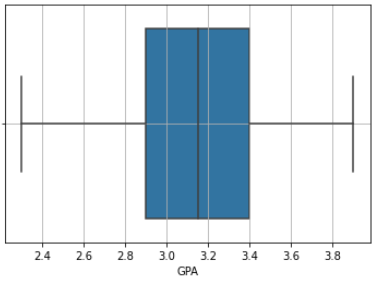
***Plot 2.8.1***

#### 2 Peaks

* **Left Skewed**
* It doesn’t Show a perfect ‘Bell Curve
* **5 Number Summary (‘GPA)**
  + count 62.00
  + mean 3.13
  + std 0.38
  + min 2.30
  + 25% 2.90
  + 50% 3.15
  + 75% 3.40
  + max 3.90
* Distribution is **Left Skewed** –
  + (Q1 – Min) which is 0.60 > Max – Q3 which is 0.50

##### Following points support the symmetry –

* + (Q2 – Min ) which is 0.75 = (Max – Q2) which is 0.75
  + (Q2 – Q1 ) which is 0.25 = (Q3 – Q2 ) which is 0.25
* **Boxplot (GPA)**



***Plot 2.8.2***

* There are No outliers
* Distribution is **Left Skewed** due to longer Left Whisker
* **Empirical Rule (GPA)**

|  |  |  |
| --- | --- | --- |
|  | **Expected Count** | **Actual Count** |
| Mean + 1 Sigma | 21.08 | 24 |
| Mean – 1 Sigma | 21.08 | 21 |
| Mean + 2 Sigma | 29.45 | 30 |
| Mean – 2 Sigma | 29.45 | 30 |
| Mean + 3 Sigma | 30.907 | 31 |
| Mean – 3 Sigma | 30.907 | 31 |

***Table 2.8.1***

#### ‘1 standard deviation’ from Mean is asymmetrical (Left Skewed)

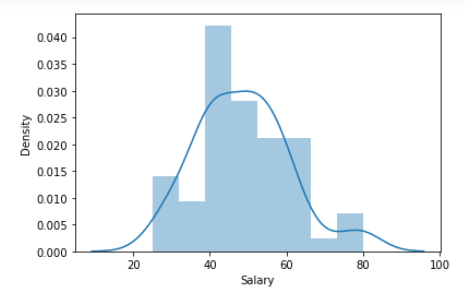
#### ‘2 standard deviation’ from Mean is approximate symmetrical

#### ‘3 standard deviation’ from Mean is approximate symmetrical

* **Mean, Median and Mode Review(GPA)**
* Mean = 3.129032258064516 Median = 3.1500000000000004 Mode = 3.1666666666666665
* Mode > Median > Mean hence the distribution is **Left Skewed**

Variable ‘Salary’ Analysis

* **HISTOGRAM(‘Salary’)**



***Plot 2.8.3***

#### 2 Peaks

#### Right Skewed

#### It doesn’t Show a perfect ‘Bell Curve’

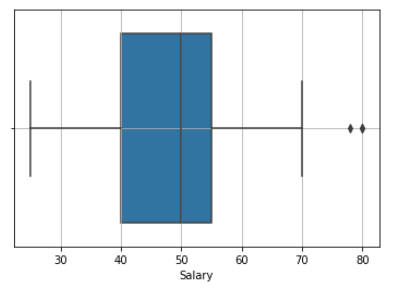
* **5 Number Summary(‘Salary’)**
  + count 62.00
  + mean 48.55
  + std 12.08
  + min 25.00
  + 25% 40.00
  + 50% 50.00
  + 75% 55.00
  + max 80.00

##### Distribution is **right skewed** –

* + (Q1 – Min) which is 0.15 < (Max – Q3) which is 0.25
  + (Q2 – Min) which is 0.25 <(Max – Q2) which is 0.30

##### Distribution is Left skewed –

* + (Q2 – Q1) which is 0.10 > (Q3 – Q2) which is 0.05
* **Boxplot(‘Salary’)**



***Plot 2.8.4***

##### There are multiple outliers

##### Distribution appears to be Asymmetric

##### Distribution is definitely **Right skewed**.

* **Empirical Rule(‘Salary’)**

|  |  |  |
| --- | --- | --- |
|  | **Expected Count** | **Actual Count** |
| Mean + 1 Sigma | 21.08 | 27 |
| Mean – 1 Sigma | 21.08 | 22 |
| Mean + 2 Sigma | 29.45 | 29 |
| Mean – 2 Sigma | 29.45 | 30 |
| Mean + 3 Sigma | 30.907 | 32 |
| Mean – 3 Sigma | 30.907 | 30 |

***Table 2.8.2***

##### ‘1 standard deviation’ from Mean is Highly Left Skewed asymmetrical which is compensated covered by 3SD

##### ‘2 standard deviation’ from Mean is approximate symmetrical

##### ‘3 standard deviation’ from Mean is approximate symmetrical

* **Mean, Median and Mode Review (‘Salary’)**

##### Mean = 48.54838709677419 Median = 50.0 Mode = 40.00

##### Skew orientation cannot be conclude however this is not a normal distribution

Variable ‘Spending’ Analysis

* **HISTOGRAM (‘Spending’)**

#### 

***Plot 2.8.5***

#### Peaks

#### Right Skewed

#### It doesn’t Show a ‘Bell Curve’

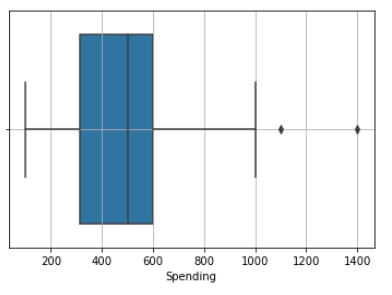
* **5 Number Summary (‘Spending’)**
  + count 62.00
  + mean 482.02
  + std 221.95
  + min 100.00
  + 25% 312.50
  + 50% 500.00
  + 75% 600.00
  + max 1400.00

##### Distribution is right skewed –

* + (Q1 – Min) which is 212.50 < (Max – Q3) which is 800
  + (Q2 – Min) which is 400 <(Max – Q2) which is 900

##### Distribution is Left skewed –

* + (Q2 – Q1) which is 187.50 > (Q3 – Q2) which is 100
* **Boxplot(‘Spending’)**



***Plot 2.8.6***

##### There are multiple outliers

##### Distribution appears to be Asymmetric

##### Distribution is Right skewed due to outliers and whiskers.

* **Empirical Rule(‘Spending’)**

|  |  |  |
| --- | --- | --- |
|  | **Expected Count** | **Actual Count** |
| Mean + 1 Sigma | 21.08 | 30 |
| Mean – 1 Sigma | 21.08 | 20 |
| Mean + 2 Sigma | 29.45 | 31 |
| Mean – 2 Sigma | 29.45 | 28 |
| Mean + 3 Sigma | 30.907 | 33 |
| Mean – 3 Sigma | 30.907 | 28 |

***Table 2.8.3***

##### ‘1 standard deviation’ from Mean is asymmetrical (Right Skewed)

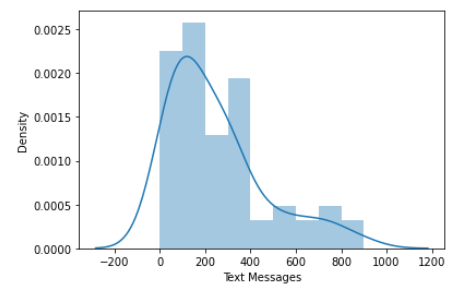
##### ‘2 standard deviation’ from Mean is asymmetrical (Right Skewed)

##### ‘3 standard deviation’ from Mean is asymmetrical (Right Skewed)

* **Mean, Median and Mode Review (‘Spending’)**
* Mean = 482.01612903225805 Median = 500.0 Mode = 0 500
* Skew orientation cannot be conclude however this is not a normal distribution

Variable ‘Text Messages’ Analysis

* **HISTOGRAM(‘Text Messages’)**



***Plot 2.8.7***

Histogram Review for CMSU[‘Text Messages’] Shows –

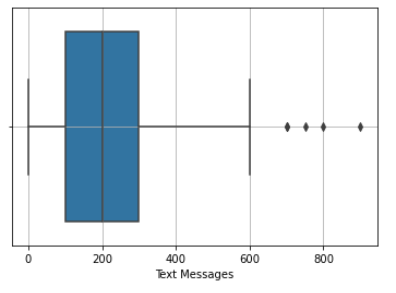
* 2 Peaks
* Right Skewed
* It doesn’t Show a ‘Bell Curve’
* **5 Number Summary(‘Text Messages’)**
  + count 62.00
  + mean 246.21
  + std 214.47
  + min 0.00
  + 25% 100.00
  + 50% 200.00
  + 75% 300.00
  + max 900.00

##### Distribution is right skewed –

* + (Q1 – Min) which is 100 < (Max – Q3) which is 600
  + (Q2 – Min) which is 200 <(Max – Q2) which is 700

##### Distribution is Symmetrical –

* + (Q2 – Q1) which is 100 = (Q3 – Q2) which is 100
* **Boxplot(‘Text Messages’)**



***Plot 2.8.8***

##### There are multiple outliers

##### Distribution appears to be Asymmetric

##### Distribution is definitely Right skewed.

* **Empirical Rule(‘Text Messages’)**

|  |  |  |
| --- | --- | --- |
|  | **Expected Count** | **Actual Count** |
| Mean + 1 Sigma | 21.08 | 16 |
| Mean – 1 Sigma | 21.08 | 33 |
| Mean + 2 Sigma | 29.45 | 21 |
| Mean – 2 Sigma | 29.45 | 36 |
| Mean + 3 Sigma | 30.907 | 25 |
| Mean – 3 Sigma | 30.907 | 36 |

***Table 2.8.4***

##### ‘1 standard deviation’ from Mean is asymmetrical (Right Skewed)

##### ‘2 standard deviation’ from Mean is asymmetrical (Right Skewed)

##### ‘3 standard deviation’ from Mean is asymmetrical (Right Skewed)

* **Mean, Median and Mode Review (‘Text Messages’)**

##### Mean = 246.20967741935485 Median = 200.0 Mode = 0 300

##### Skew orientation cannot be concluded from mean/ median/mode, however this is not a normal distribution

Central Tendency of continuous variables in question is analysed against 5 techniques / methods and it is evident that **none of these 4 variables follow a Normal Distribution**, Please refer **Table *2.8.5***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **GPA** | **Salary** | **Spending** | **Text Messages** |
| **Analysis** |  |  |  |  |
| HISTOGRAM | Left Skewed | Right skewed | Right skewed | Right skewed |
| 5 Number Summary | Left Skewed | Right skewed | Right skewed | Right skewed |
| Boxplot | Left Skewed | Right skewed | Right skewed | Right skewed |
| Empirical Rule | Left Skewed | Right Skewed | Right Skewed | Right Skewed |
| Mean, Median and Mode Review | Left Skewed | Non conclusive | Non conclusive | Non conclusive |
| CONCLUSION | Not a Normal Distribution | Not a Normal Distribution | Not a Normal Distribution | Not a Normal Distribution |

***Table 2.8.5***

### **Write a note summarizing your conclusions for this whole Problem 2.**

* 62 students participated in the survey including 33 Female and 29 Male
* None of the Female students are ‘undecided’ about the **Major** where 3 Males are yet to decide
* Proportion of male intending to Graduate is **more than** female
* There are more male students undertaking **‘Full-Time’** employment than female
* There are more Female students undertaking **‘Part-Time’** employment than female
* Proportion of female Unemployed is **more than** male

Based on the above finding a campaign should be launched within University targeting; ‘Gender’ community to spread awareness and improve education level as well as employment.

# Problem 3

An important quality characteristic used by the manufacturers of ABC asphalt shingles is the amount of moisture the shingles contain when they are packaged. Customers may feel that they have purchased a product lacking in quality if they find moisture and wet shingles inside the packaging.   In some cases, excessive moisture can cause the granules attached to the shingles for texture and colouring purposes to fall off the shingles resulting in appearance problems. To monitor the amount of moisture present, the company conducts moisture tests. A shingle is weighed and then dried. The shingle is then reweighed, and based on the amount of moisture taken out of the product, the pounds of moisture per 100 square feet is calculated. The company would like to show that the mean moisture content is less than 0.35 pound per 100 square feet.

The file ([A & B shingles.csv](https://olympus.greatlearning.in/courses/39182/files/2197038/download?verifier=BvyyRdHbmm31rN3jJWMAKzaQkqwAZgqqjPLnCoYx&wrap=1)) includes 36 measurements (in pounds per 100 square feet) for A shingles and 31 for B shingles.

## Do you think there is evidence that means moisture contents in both types of shingles are within the permissible limits? State your conclusions clearly showing all steps.

Solution to this problem is performed by following steps –

* Define Test Objectives
* Test Process and Conclusion
  + Defining H0 and HA
  + Deciding the significance level
  + Identify Test Statistics
  + Calculate P-value and Test Statistics
  + Decide to reject or accept null hypothesis

Test Objective

For every moisture test, the claim to check here becomes whether the moisture content is still greater than 0.35 pounds per 100 square feet

**# Step 1: Define null and alternative hypotheses**

For the **A shingles**, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:

* H0 : mean moisture content <=0.35
* HA : mean moisture content > 0.35

For the B shingles, the null and alternative hypothesis to test whether the population mean moisture content is less than 0.35 pound per 100 square feet is given:

* H0 : mean moisture content <=0.35
* HA : mean moisture content > 0.35

**# Step 2: Decide the significance level**

We select Significance Level ( **) = 0.05** as this is not specified in the problem statement

**# Step 3: Identify Test Statistics**

* We do not know the population standard deviation and for A, **sample size is 36**. So we use the t distribution and the $t\_{STAT}$ test statistic.
* We do not know the population standard deviation and for B, **sample size is 31**. So we use the t distribution and the $t\_{STAT}$ test statistic.

**# Step 4: Calculate P-value and Test Statistics**

* Our one-sample t-test for A shingles ; **t statistic: -1.4735046253382782 ,**  **p-value= 0.0748\***
* Our one-sample t-test for B shingles ; **t statistic: -3.1003313069986995** , **p-value= 0.0021\***

\*- Just for information, One Sample T – test by default returns 2 tail results, in this case we have to perform 1 tail test only and therefore p-Value is divided by two.

**# Step 5: Decide to reject or accept null hypothesis**

#### FOR 'A' shingles

We have no evidence to reject the null hypothesis for A shingles; since p value for A shingles > Level of significance Our one-sample t-test for A shingles; p-value= 0.0748

Basis the hypothesis test performed for the given Sample of 36 observations at 95% confidence level **we fail to reject H0** i.e. average moisture content For 'A' Shingles is less than 0.35 pounds per 100 square feet

#### FOR 'B' shingles

We have evidence to reject the null hypothesis for B shingles ; since p value for B shingles < Level of significance Our one-sample t-test for B shingles ;p-value= 0.0021

Basis the hypothesis test performed for the given Sample of 31 observations at 95% confidence level **we reject H0** i.e. average moisture content for 'B' Shingles is more than 0.35 pounds per 100 square feet

## Do you think that the population mean for shingles A and B are equal? Form the hypothesis and conduct the test of the hypothesis. What assumption do you need to check before the test for equality of means is performed?

Solution to this problem is performed by following steps –

* Define Test Objectives
* Assumptions Review
* Test Process and Conclusion
  + Defining H0 and HA
  + Deciding the significance level
  + Identify Test Statistics
  + Calculate P-value and Test Statistics
  + Decide to reject or accept null hypothesis

Test Objective

#### To identify if population mean for shingles A and B are equal. Essentially to compare the quality of both the sets ‘A’ and ’B’ in terms of moisture contents.

Assumptions

* Distribution of the two populations is **normal**.
* Variance of two sets being compared is same; in this situation variance of both the data set has a negligible difference of **0.0005**

Test Process and Conclusion

**# Step 1: Define null and alternative hypotheses**

In testing whether the average moisture content of the shingles is same in both the categories **('A' and 'B') of the shingles**, the null hypothesis states that the average moisture content of the shingles are the same, *𝜇𝐴*

equals *𝜇𝐵*. The alternative hypothesis states that the average moisture content of the shingles are different, *𝜇𝐴* is not equal to *𝜇𝐵*

.

* *𝐻*0: *𝜇𝐴* - *𝜇𝐵* = 0 i.e *𝜇𝐴* = *𝜇𝐵*
* *𝐻𝐴: 𝜇𝐴 - 𝜇𝐵 ≠ 0 i.e 𝜇𝐴 ≠ 𝜇𝐵*

**# Step 2: Decide the significance level**

* We select Significance Level (**𝛼 ) = 0.05** as this is not specified in the problem statement
* Population standard deviation is not known.

**# Step 3: Identify Test Statistics**

* We have two samples and we do not know the population standard deviation.
* Sample sizes for both samples are not the same.
* We use the t distribution and the, test statistic for **2sampleT-Test(ttest\_ind)** test.

**# Step 4: Calculate P-value and Test Statistics**

* We use the **scipy.stats.ttest\_ind** to calculate the t-test for the means of TWO INDEPENDENT samples of scores given the two sample observations. This function returns t statistic and two-tailed p value.
* This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances.
* For this exercise, we are going to first assume that the variance is equal and then compute the necessary statistical values.
* two-sample t-test p-value; tstat 1.2896282719661123, p-value= 0.2017

**# Step 5: Decide to reject or accept null hypothesis**

**In this situation p-Value** , 0.2017 > 0.05

Hence, we **do not have enough evidence to reject the null hypothesis** in favour of alternative hypothesis. We conclude that the average moisture content of the shingles in both the categories A and B ARE same.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End of Document \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*